Policy Gradient Algorithms Series

ritchie huang

2019-09-18

Contents

1 Preliminaries

From the paper, **TRPO** algorithm finally propose the core optimization problem:

$$\begin{aligned} \theta_{k+1} &= \underset{\theta}{argmax}\, \mathcal{L}(\theta_k, \theta) \\ s.t. \quad & \frac{1}{2}(\theta - \theta_k)^T H(\theta - \theta_k) \leq \delta \end{aligned} \tag{1}$$

2 First order approximation on the gradient

In trust region policy optimization, we finally get this:

$$\mathcal{L}(\theta_k, \theta) = \underset{s,a \in \pi_{\theta_k}}{\mathbb{E}} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_k(a|s)}} A^{\pi_{\theta_k}}(s, a) \right] \tag{2}$$

if we perform first-order Tylor expansion around the old policy θ_k on this equation, we obtain:

$$\begin{aligned} \mathcal{L}(\theta_k, \theta) &= \mathcal{L}(\theta_k, \theta_k) + \nabla_\theta \mathcal{L}(\theta_k, \theta)|_{\theta=\theta_k}(\theta - \theta_k) \\ &= 0 + \nabla_\theta \mathcal{L}(\theta_k, \theta)|_{\theta=\theta_k}(\theta - \theta_k) \\ &= \nabla_\theta \mathcal{L}(\theta_k, \theta)|_{\theta=\theta_k} \end{aligned} \tag{3}$$

we take the gradient w.r.t θ out:

$$\begin{aligned} \nabla_\theta \mathcal{L}(\theta_k, \theta)|_{\theta=\theta_k} &= \underset{s,a \in \pi_{\theta_k}}{\mathbb{E}} \left[\frac{\nabla_\theta \pi_\theta(a|s)|_{\theta=\theta_k}}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s, a) \right] \\ &= \underset{s,a \in \pi_{\theta_k}}{\mathbb{E}} [\nabla_\theta \log \pi_\theta(a|s)|_{\theta=\theta_k} A^{\pi_{\theta_k}}(s, a)] \end{aligned} \tag{4}$$

Coincidently, the gradient is exactly the gradient of "Vanilla Policy Gradient".

3 Second order approximation on the constraint

$$\begin{aligned} Pr\left[\hat{Z} > (1+\epsilon)z \quad or \quad \hat{Z} < (1-\epsilon)z\right] &= Pr\left[|\hat{Z} - z| > \epsilon z\right] \\ &= Pr\left[|\frac{1}{\bar{Y}} - 1 - z| > \epsilon z\right] \end{aligned} \tag{5}$$